Speech Systems - Course Introduction

K Sri Rama Murty

IIT Hyderabad

ksrm@ee.iith.ac.in

August 4, 2022

Why Speech?

- Speech is the most natural forms of communication
  - Does not require any special training

Why Speech?

- Speech is the most natural forms of communication
  - Does not require any special training
- Flexible
  - Hands/Eyes free communication
  - Radio survived the era of HD television!
  - Assistive technologies for blind
  - Interaction while driving - Google maps instructions

Why Speech?

- Speech is the most natural forms of communication
  - Does not require any special training
- Flexible
  - Hands/Eyes free communication
  - Radio survived the era of HD television!
  - Assistive technologies for blind
  - Interaction while driving - Google maps instructions
- Efficient
  - Conveys lot more information than mere text content

Why Speech?

- Speech is the most natural forms of communication
  - Does not require any special training
- Flexible
  - Hands/Eyes free communication
  - Radio survived the era of HD television!
  - Assistive technologies for blind
  - Interaction while driving - Google maps instructions
- Efficient
  - Conveys lot more information than mere text content
- Economical
  - Inexpensive transmission and reception of information
  - Voice communication is the reason behind success of mobile phones

Information in Speech Signal

- Linguistic Information
  - Information that can be represented by a set of discrete symbols
  - Conveys textual message in the speech signal
  - Language recognition, speech recognition, search for specific words

Information in Speech Signal

- Linguistic Information
  - Information that can be represented by a set of discrete symbols
  - Conveys textual message in the speech signal
  - Language recognition, speech recognition, search for specific words
- Paralinguisitc Information
  - Supplemental information that is not inferable from written text
  - Added by speaker to modify & supplement the linguistic information
  - Intention, attitude, emphasis, speaking style, etc.
  - It can be consciously controlled by the speaker

Information in Speech Signal

- Linguistic Information
  - Information that can be represented by a set of discrete symbols
  - Conveys textual message in the speech signal
  - Language recognition, speech recognition, search for specific words
- Paralinguisitc Information
  - Supplemental information that is not inferable from written text
  - Added by speaker to modify & supplement the linguistic information
  - Intention, attitude, emphasis, speaking style, etc.
  - It can be consciously controlled by the speaker
- Nonlinguistic Information
  - Factors that cannot, generally, be controlled by the speaker
  - Gender, speaker identity, age, physical build, emotional state, health, idiosyncrasy (filler words, mother-tongue), etc.

Information in Speech Signal

- Linguistic Information
  - Information that can be represented by a set of discrete symbols
  - Conveys textual message in the speech signal
  - Language recognition, speech recognition, search for specific words
- Paralinguisitc Information
  - Supplemental information that is not inferable from written text
  - Added by speaker to modify & supplement the linguistic information
  - Intention, attitude, emphasis, speaking style, etc.
  - It can be consciously controlled by the speaker
- Nonlinguistic Information
  - Factors that cannot, generally, be controlled by the speaker
  - Gender, speaker identity, age, physical build, emotional state, health, idiosyncrasy (filler words, mother-tongue), etc.
- Background Information - Acoustic environment around the speaker

Speech Studies - Highly Interdisciplinary Subject

- Linguistics - The scientific study of language
  - Includes phonetics, phonology, morphology, syntax, semantics

Speech Studies - Highly Interdisciplinary Subject

- Linguistics - The scientific study of language
  - Includes phonetics, phonology, morphology, syntax, semantics
- Physiology of speech production
  - Anatomy of articulators, health-related cues

Speech Studies - Highly Interdisciplinary Subject

- Linguistics - The scientific study of language
  - Includes phonetics, phonology, morphology, syntax, semantics
- Physiology of speech production
  - Anatomy of articulators, health-related cues
- Physics - Acoustics of speech production & propagation

Speech Studies - Highly Interdisciplinary Subject

- Linguistics - The scientific study of language
  - Includes phonetics, phonology, morphology, syntax, semantics
- Physiology of speech production
  - Anatomy of articulators, health-related cues
- Physics - Acoustics of speech production & propagation
- Psychology - Assessing mental state of a person

Speech Studies - Highly Interdisciplinary Subject

- Linguistics - The scientific study of language
  - Includes phonetics, phonology, morphology, syntax, semantics
- Physiology of speech production
  - Anatomy of articulators, health-related cues
- Physics - Acoustics of speech production & propagation
- Psychology - Assessing mental state of a person
- Communication Systems

Speech Studies - Highly Interdisciplinary Subject

- Linguistics - The scientific study of language
  - Includes phonetics, phonology, morphology, syntax, semantics
- Physiology of speech production
  - Anatomy of articulators, health-related cues
- Physics - Acoustics of speech production & propagation
- Psychology - Assessing mental state of a person
- Communication Systems
  - Coding and transmission of speech signals

Speech Studies - Highly Interdisciplinary Subject

- Linguistics - The scientific study of language
  - Includes phonetics, phonology, morphology, syntax, semantics
- Physiology of speech production
  - Anatomy of articulators, health-related cues
- Physics - Acoustics of speech production & propagation
- Psychology - Assessing mental state of a person
- Communication Systems
  - Coding and transmission of speech signals
- Signal processing
  - Signal detection, enhancement and feature extraction
- AI & Machine Learning
  - Data-driven models for information extraction
  - Speech recognition, speech synthesis, speaker recognition, language recognition, etc.

Probabilistic Approach towards Speech Systems

- Speech signal contains several sources of variability
  - W - Textual message
  - S - Speaker identity
  - L - Language of communication
  - E - Emotional state of the speaker
  - B - Ambient background and recording conditions

Probabilistic Approach towards Speech Systems

- Speech signal contains several sources of variability
  - W - Textual message
  - S - Speaker identity
  - L - Language of communication
  - E - Emotional state of the speaker
  - B - Ambient background and recording conditions
  - Observed signal $o[t]$ is a sample function drawn from $p(w, s, l, e, v, b)$

Probabilistic Approach towards Speech Systems

- Speech signal contains several sources of variability
  - W - Textual message
  - S - Speaker identity
  - L - Language of communication
  - E - Emotional state of the speaker
  - B - Ambient background and recording conditions
  - Observed signal $o[t]$ is a sample function drawn from $p(w, s, l, e, v, b)$
- Speaker recognition - Marginalize all other factors except S

$$p(s/o[t]) = \sum_{W,L,E,B} p(w, s, l, e, b/o[t])$$

Probabilistic Approach towards Speech Systems

- Speech signal contains several sources of variability
  - W - Textual message
  - S - Speaker identity
  - L - Language of communication
  - E - Emotional state of the speaker
  - B - Ambient background and recording conditions
  - Observed signal $o[t]$ is a sample function drawn from $p(w, s, l, e, v, b)$
- Speaker recognition - Marginalize all other factors except S

$$p(s/o[t]) = \sum_{W,L,E,B} p(w, s, l, e, b/o[t])$$

- Speech recognition - Marginalize all other factors except W

$$p(w/o[t]) = \sum_{S,L,E,B} p(w, s, l, e, b/o[t])$$

Course Outline - Basics

- Acoustic theory of speech production (Domain knowledge)
  - Study of speech sounds through analysis of waveform and spectrum
  - Characteristics of vowels, stop-consonants and fricatives

Course Outline - Basics

- Acoustic theory of speech production (Domain knowledge)
  - Study of speech sounds through analysis of waveform and spectrum
  - Characteristics of vowels, stop-consonants and fricatives
- Feature extraction (Signal processing)
  - Compact representation of desired information in the speech signal
  - Speech is a nonstationary signal: short-term spectral analysis

Course Outline - Basics

- Acoustic theory of speech production (Domain knowledge)
  - Study of speech sounds through analysis of waveform and spectrum
  - Characteristics of vowels, stop-consonants and fricatives
- Feature extraction (Signal processing)
  - Compact representation of desired information in the speech signal
  - Speech is a nonstationary signal: short-term spectral analysis
- Building speech systems
  - Probability density estimation - Gaussian mixture models
  - Extracting utterance-level representation from speech signal
  - Speaker recognition, language recognition and emotion recognition

Course Outline - Speech Recognition

- Acoustic modeling - P(feature/word)
  - Estimate likelihood of a sound unit from features
  - pdf estimation using hidden Markov models (piece-wise stationary RP)

Course Outline - Speech Recognition

- Acoustic modeling - P(feature/word)
  - Estimate likelihood of a sound unit from features
  - pdf estimation using hidden Markov models (piece-wise stationary RP)
- Language modeling - P(word)
  - Estimate prior probability of a sound unit. P(I/am) vs P(eye/am)?
  - Imposing syntactic constraints of the language

Course Outline - Speech Recognition

- Acoustic modeling - P(feature/word)
  - Estimate likelihood of a sound unit from features
  - pdf estimation using hidden Markov models (piece-wise stationary RP)
- Language modeling - P(word)
  - Estimate prior probability of a sound unit. P(I/am) vs P(eye/am)?
  - Imposing syntactic constraints of the language
- Acoustic decoding - P(word/features)
  - Estimate posterior from prior and likelihood
  - Decode the best possible word sequence from the feature sequence

Course Outline - Speech Recognition

- Acoustic modeling - P(feature/word)
  - Estimate likelihood of a sound unit from features
  - pdf estimation using hidden Markov models (piece-wise stationary RP)
- Language modeling - P(word)
  - Estimate prior probability of a sound unit. P(I/am) vs P(eye/am)?
  - Imposing syntactic constraints of the language
- Acoustic decoding - P(word/features)
  - Estimate posterior from prior and likelihood
  - Decode the best possible word sequence from the feature sequence
- Deep acoustic modeling
  - Discriminative models for likelihood estimation using DNNs (pretend)
  - End-to-end models for ASR (waveform to character)

If Time Permits...

- Towards unsupervised speech recognition
  - Self supervised feature extraction
  - Leveraging large amounts of non-parallel text and speech corpora

If Time Permits...

- Towards unsupervised speech recognition
  - Self supervised feature extraction
  - Leveraging large amounts of non-parallel text and speech corpora
- Speech synthesis
  - Estimate joint density from marginals.
  - Ill posed problem - Difficult to infer dependencies
  - Draw a sample function from the estimated joint density

Suggested Prerequisites

- Basics of signal processing
- Probability and random processes
- Foundations of machine learning
  - Supervised learning, regression, classification
- Neural networks
  - Understanding of forward and back propagation equations
  - Basic exposure to deep learning
- Proficiency in Python

Evaluation Criteria

- Homework & Programming Assignments - 40%
- Class tests - 15%
- Project & viva-voce - 25%
- Endexam - 20%

Evaluation Criteria

- Homework & Programming Assignments - 40%
- Class tests - 15%
- Project & viva-voce - 25%
- Endexam - 20%
- Programming assignments have to be done in Python.
- Project theme has to be finalized by the end of the first segment
- Handwritten summary of lecture has to be submitted on the same day
- Delayed submissions will be penalized

Thank You!